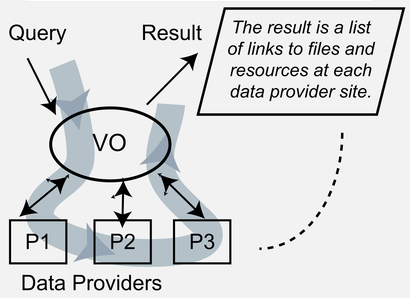
Data Unification and Analysis Services for Magnetospheric and Heliospheric Data

Jon Vandegriff, Johns Hopkins University Applied Physics Lab jon.vandegriff@jhuapl.edu

http://sd-www.jhuapl.edu/MIDL

THE ISSUE

The small box approach is not enough for the timeseries style datasets typical of the in-situ magnetospheric and heliospheric measurements.



For image data, there is a limited number of file types that can be returned, and so the task of viewing and analyzing the VO results is possible.

But for in-situ data, the files discovered will be in many different formats - HDF5, HDF4, CDF, netCDF, ASCII and custom binary formats.

More importantly, the inner arrangement of files in the same format can differ dramatically. Two datasets in ASCII can still have very different layouts - one could be CSV (comma separated) with the date in Year, Month, Day, and other ASCII files could have different formats.

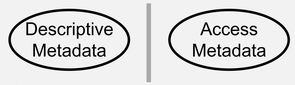
For files in CDF, some standards exists, but as datasets become increasingly distributed, there will be more and more data with "native" (i.e., non-standard formats).

Format translation is not enough - a mechanism is needed to understand the semantics of the data being accessed.

AN EXTENSION TO THE SMALL BOX

In order to keep the data model clean and simple, science data types should be described in semantic terms using a data interface, which provides an API to the expected science content of a dataset.

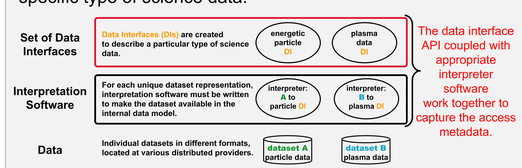
There needs to be a clear separation in the data model between characteristics of a dataset (measurement type, overall time range, mission name, spectral range, etc) and the way to access the data.



In the OPeNDAP world, these are called "search" and "use" metadata respectively. An OPeNDAP paper covering the subject is attached below.

The traditional way to specify "access metadata" has been in the same way as "descriptive metadata." But a more efficient way to capture the access metadata is to use a totally different mechanism which more naturally captures the "hows" of accessing the expected content in a dataset.

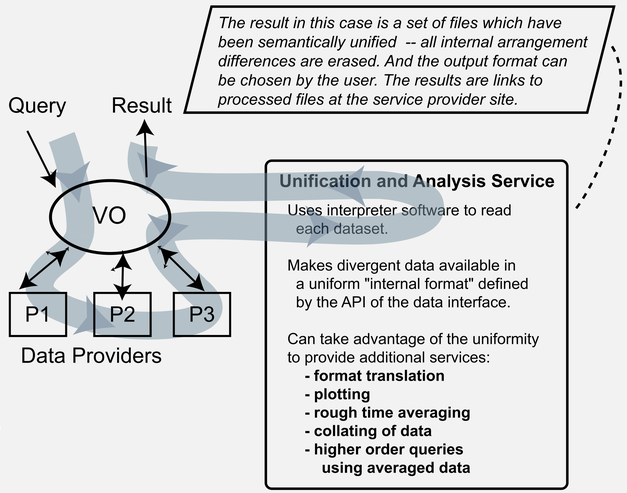
The interface idea behind web services (i.e., a well defined interface to a well advertised service) can be extended to create the concept of a "data interface." A data interface is an API defining semantically what can be extracted from a specific type of science data.



A data interface in fact hides the access details, delegating it to another place outside the data model. The actual access takes place in "interpreter software," which reads the data in its native format and fulfills the contract of the data interface.

One piece of interpreter software needs to be written for each dataset to be included in such a system. But for similar datasets, interpreter software can be shared. Generic interpreter software could be written for some datasets. The settings used to drive the generic interpreter software would then be the place to store access metadata.

The semantic interpretation of data can be added to the VO architecture as a service.



Sample Data Interfaces:

Energetic Particles

```
public interface IParticleData
{
    public static String ID_LABEL = "Particle Data";
    public int getNumChannels();
    public IParticleChannelDescription getChannel(int i);
    public double [] getTimeStamp(IParticleChannel d);
    public double [] getAccumulationTime(IParticleChannel d);
    public double [] getIntensity(IParticleChannel d);
    public double [] getUncertainty(IParticleChannel d);
}
```

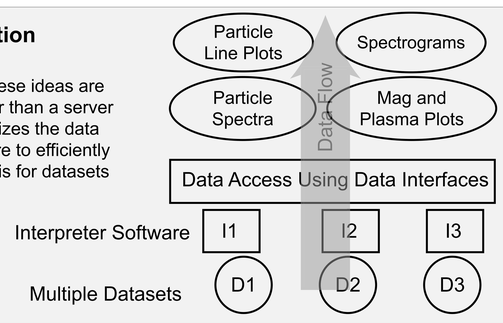
Magnetic Field

```
public interface IMagneticField
{
    public static String ID_LABEL = "Magnetic Field Data";
    public double[] getTimeStamp();
    public double[] getBx();
    public double[] getBy();
    public double[] getBz();
    public double[] getBxUncertainty();
    public double[] getByUncertainty();
    public double[] getBzUncertainty();
}
```

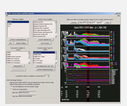
Client-Based Implementation

The current implementation of these ideas are being used in a client-side (rather than a server side) analysis program which utilizes the data interfaces and interpreter software to efficiently provide data retrieval and analysis for datasets from 7 different missions:

Cassini
ACE, IMP-8
Geotail, AMPTE, ISEE-1, ISEE-2

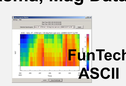


MIDL Main Screen



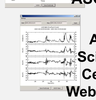
From here you choose the mission, the instrument, and the time range. Then you can launch one of several different types of plots.

ACE Particles, Plasma, Mag Data



FunTech ASCII

ACE Science Center Web Pages



Geotail Particles

APL binary

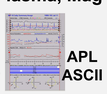
APL binary

Cassini Particles, Mag

APL binary

APL binary

IMP-8 Particles, Plasma, Mag



APL ASCII

APL ASCII

Goddard ASCII

Note that data is loaded from multiple, remote locations in various formats.